Acknowledgement

First of all, we would like to express our heartfelt gratitude to our project supervisor Mr. Dhiraj Shrestha for his support during the project and co-operating with us to help us carry our project smoothly. His experiences in the field have been a great asset for our project. His encouraging words and working techniques have made this project a successful one.

Our deepest appreciation to all those who provided us the possibility to complete this project. We extend our gratitude to all who have deliberately or unknowingly added a brick in the completion of this project. We enjoyed the duration of the work studying different modules and creating this project.

Taking this opportunity, we would like to thank all our peers and classmates who directly or indirectly helped us in making this project a successful one be it by encouraging us throughout the project or else through their valuable suggestions which we have tried our best to assimilate within our work.

Abstract

Disease Risk Prediction is the computational method to predict disease risk. Our project entitled "Self-diagnosis" is machine learning based project that helps users to predict the risk of the disease by analyzing the available data set given to the machine. It predicts how much the disease is likely to occur analyzing the previously recorded diagnosis. The main aim of this project is to help the patients analyze the disease forehand and take necessary preventions before consulting any medical professionals. This project is accomplished with the use of Naïve Bayes Algorithm for pattern recognition and prediction using python programming language and tools like Weka for data analysis and comparison, Django framework for user interface and Jupyter notebook as text editor and MATLAB for data visualization.

Keywords: *data,data-preprocessing, Naïve Bayes, computer aided diagnosis, data visualisation*

Table of Contents

Title Page

List of Abbreviation

ML - Machine Learning

WHO - World Health Organization

ADPS - Automated Disease Prediction System

IDF - International Diabetes Foundation

SEA - South East Asia

CAD - Computer Aided Design

GDPS - General Disease Prediction System

CSV - Comma-Separated Values

WEKA - Waikato Environment for Knowledge Analysis

TP - True Positives

TN – True Negatives

FP – False Positives

FN - False Negatives

List of Figure

List of Table

Chapter 1: Introduction

1.1 Background

The exponential increase in medical data collection and management has drawn more attention in disease predictions from perception of big data inquiry. Numerous experiments and strategies have been conducted to classify and improve the condition of disease recognition involving a large number of medical personals. The majority of these wellness programs include an annual screening to detect individuals with the highest risk of developing a chronic disease. However, the percentage of successful recognition in a curable time is very low. A very significant number of patients, despite receiving a proper treatment, die due to the delay. How can big data analysis expertise be used to recognize the disease and generate a better diagnosis method?

These problems of disease prediction based on the symptoms can be solved with the following strategies - first the system will use Decision tree mapping algorithm to generate the pattern and causes of disease. It clearly shows the diseases and sub diseases. Second, we increase the operational efficiency by using an algorithm to divide the data into partitions.

1.2 Objectives

The objectives of this project are:

- Predicting diseases and sub diseases with 75% to 80% accuracy based on the preliminary symptoms from patients.
- Guessing the risk of obtaining Hepatitis and Diabetes by Computer Aided Diagnosis
- To get familiar with Data Analysis and Machine Learning using Python Development Environment and its libraries.

1.3 Motivation and Significance

Number of internet users is growing exponentially over the years. It’s unusual for a people of this generation to not carry a smartphone, and not spend some part of their time in the internet. People search for post their health related queries (such as asking about what kind of disease that they might be suffering from) on various healthcare forums. These predictions may not be always accurate, and also there is no assurance that users will always get a reply on their post. Moreover, some posts are fabricated or made up which can drive the patient in a wrong direction. (Samiti, 2016) According to WHO, at least 8.8 million people die because of late diagnosis of cancer every year. Similarly, there is no exact data of patients with diabetes in Nepal. 2016 Diabetes Profile has shown that 9.1 percent Nepali population are living with diabetes. It includes 10.5 percent men and 7.9 percent women. (IDF SEA members, 2018) In 2017, 4% of total adult population i.e. 657,200 cases of diabetes. Diabetes is the ninth leading cause of death in women globally, causing 2.1 million deaths per year. Up to 70% of cases of type 2 diabetes could be prevented through the adoption of a healthy lifestyle. Rapid proliferation of Internet technology and handheld devices has opened up new avenues for online healthcare system. There are instances where online medical help or healthcare advice is easier or faster to grasp than real world help. People often feel reluctant to go to hospital or physician on minor symptoms. However, in many cases, these minor symptoms may trigger major health hazards. As online health advice is easily reachable, it can be a great head start for users. Computer Aided Diagnosis (CAD) based on the history records and pattern recognition of past patients will provide the user with tentative idea of what the person is suffering from with a percent accuracy. CAD will be a huge help for medical personals like doctors in diagnosis as the pattern of disease in that region will yield a certain red flags and speed up the process of diagnosis and eventually treatment. Thus, we have been motivated to design a disease prediction system that can help people detect the disease and its risk factor.

1.4 Features

- Disease prediction is based upon the underlying symptoms and medical test results of the patients.
- User will be able to list out their preliminary symptoms and tests for a specific disease and know if they might be suffering from that certain disease or not .

Chapter 2: Related Works

Heart Disease Prediction System using Naive Bayes (Arun.R, 2018)

This paper focuses importance of quality diagnosis of patients and viable medications. It conveys on how Naive Bayes Algorithm can be used in anticipating the probability of a patient getting a disease. It has explained how mix of clinical choice help with PC based patient records could diminish medical blunders, upgrade tolerant security, diminish undesirable practice variety, and enhance understanding result proposal. It has used Advanced Encryption Standard (AES) as an encrypting algorithm for securing sensitive but unclassified material. The fundamental goal of this journal is to model Heart Disease Expectation System (HDPS) utilizing three information mining displaying strategies, Decision Trees, Naïve Bayes and Neural Network. This paper also designed a cloud assisted privacy preserving mobile health monitoring system called CAM which can effectively protect the privacy of clients and the intellectual property of Health service providers.

Automated Disease Prediction System (ADPS) (Rashid, 2016)

ADPS is a system that relies on guided user input and provides a list of topmost diseases and greater likelihood of occurrence disease. ADPS proposes RA data structure where five relevant parameters from user input i.e. Symptom name, Time, Intensity, Organ Name, and Duration are taken in account and matched with Disease Symptom Database to find the match. Word Tagging, Synonym Parent Tree and Decision Tree is made based on which probability of the disease is found and accuracy comparison process takes place. The accuracy if ADPS resembles the quality of a predicted disease as is an average of 14.35% more in comparison with existing solution with a minimum of 4.60% and maximum of 25.97%. The probability computation is done when symptoms from data matrix are retrieved and mapped with the symptoms in the database. Corresponding diseases are recorded, asymmetric binary similarity factor is calculated among the user query data matrix and matched data matrix/matrices and probability of occurrence of disease is predicted.

General Disease Prediction System (GDPS) (Shratik J. Mishra, 2018)

GDPS is a system that focuses on identifying or predicting the disease at the earliest phase to avoid any unwanted casualties. It applies data mining techniques and ID3 decision tree algorithms. This symptom will predict the most possible disease base on the given symptoms and precautionary measures required to avoid the aggression of disease. This system carries out data mining in its preliminary stages then system is trained using machine learning and data mining to predict the disease based on the input data given by the user. ID3 algorithm is used to generate a decision tree from a given dataset. ID3 works mainly on three things, firstly the entropy of each attribute, second information gain and third, entropy of whole dataset.

Disease Predictor: A Disease Prediction App (Bharat Gandhi, 2017)

The Disease Predictor app helps user to diagnose a disease in real time by selecting the various symptoms through a given list. The symptoms selected are then processed to take out the chances of a disease to occur. There can be more than one disease predicted for a same set of symptoms but may be with different percentage of chances of occurrence. The details of the disease are also given with useful information such as all the possible symptoms, detailed explanation of the disease and the next steps that are to be taken forward so as to prevent the disease without even going to any doctor or dispensary. The Disease Predictor app also creates an alert in regular time intervals after a certain prediction to ensure that whether the user has followed the steps provided post the diagnosis for ceasing the disease.

Disease Prediction Using Machine Learning Over Big Data (Vinitha S, 2018)

This journal helps to solve risk organization based on big data analysis, it see the structured and unstructured data in healthcare field to assess the risk of disease. First, the system use Decision tree map algorithm to generate the pattern and causes of disease. It clearly shows the diseases and sub diseases. Second, by using Map Reduce algorithm for partitioning the data such that a query will be analyzed only in a specific partition, which will increase the operational efficiency but reduce query retrieval time. Map reducing algorithm is used for partitioning the medical data to increase the operational efficiency based on the output of Decision Tree map. It is cost affordable. It member Map Reduce algorithm for subdividing the data such that a request would be scrutinized only in the explicit partition, which will increase effective proficiency but cut query rescue time. In tally to that, it provide definite rations for specific clients to pattern his/her condition.

Chapter 3: Procedures and Methods

3.1 System Specifications

3.1.1 Software Specifications

- Programming language: Python3, Django-web framework
- Operating System: Windows
- Tools: Anaconda Prompt, Jupyter Notebook, Weka tool

3.1.2 Hardware Specifications

- Processor: Intel i3-5th Generation processor
- RAM: 4GB, 8GB recommended

3.2 System Diagrams

3.2.1 Block Diagram

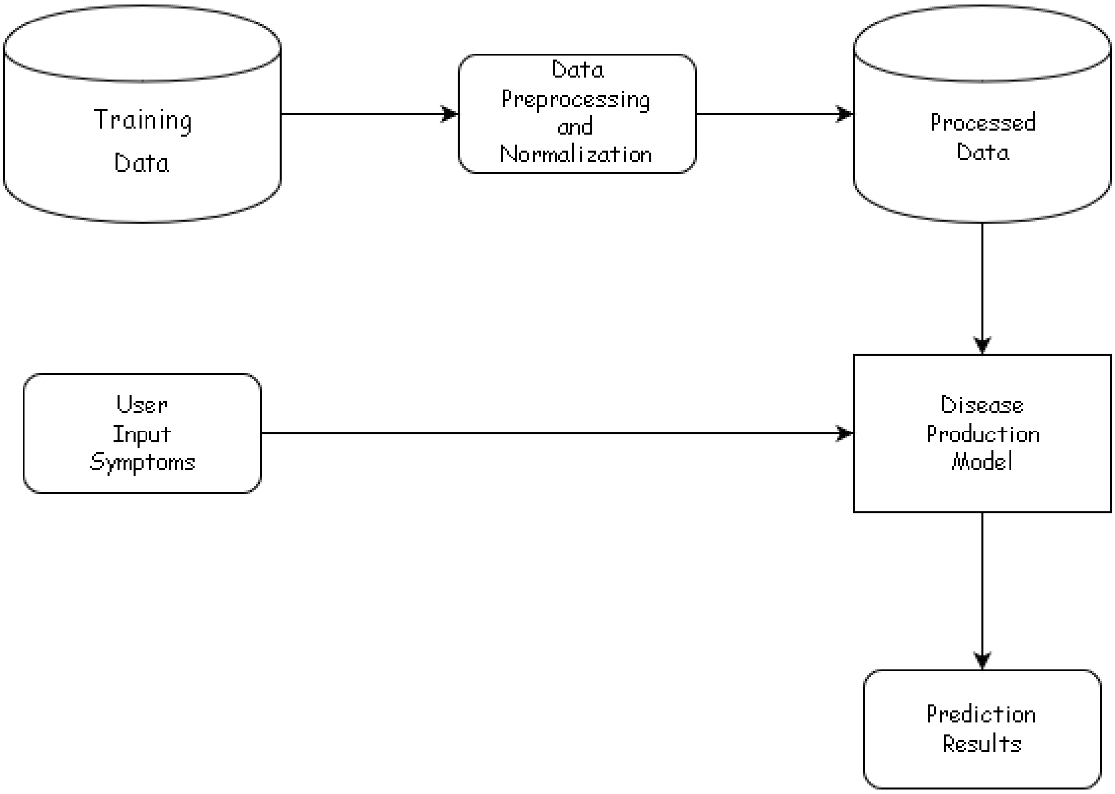


Figure 3.1.1: Block Diagram

3.2.2 Use Case Diagram

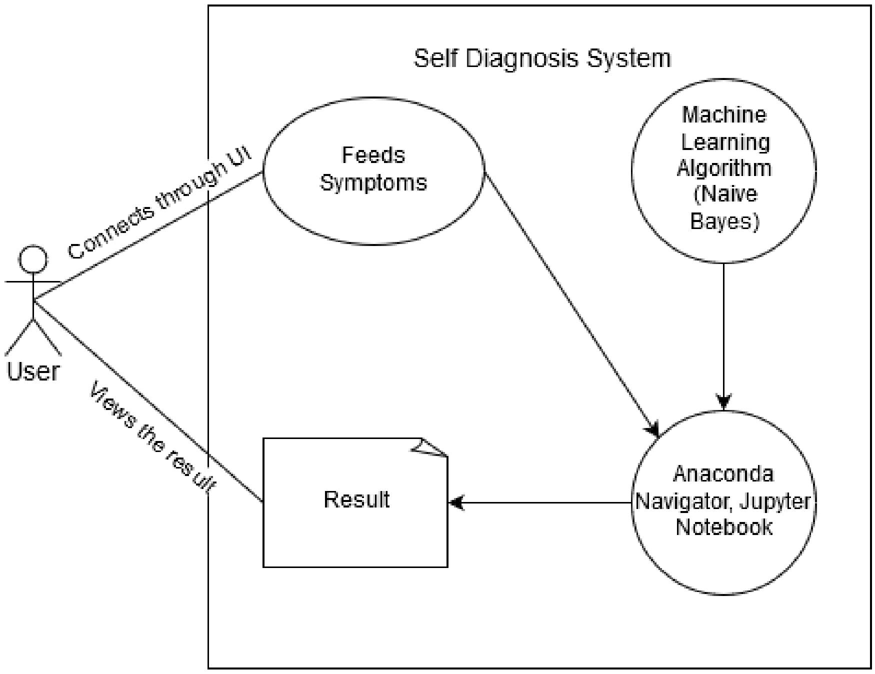


Figure 3.2.2: Use Case Diagram

3.2.3 Flow Chart

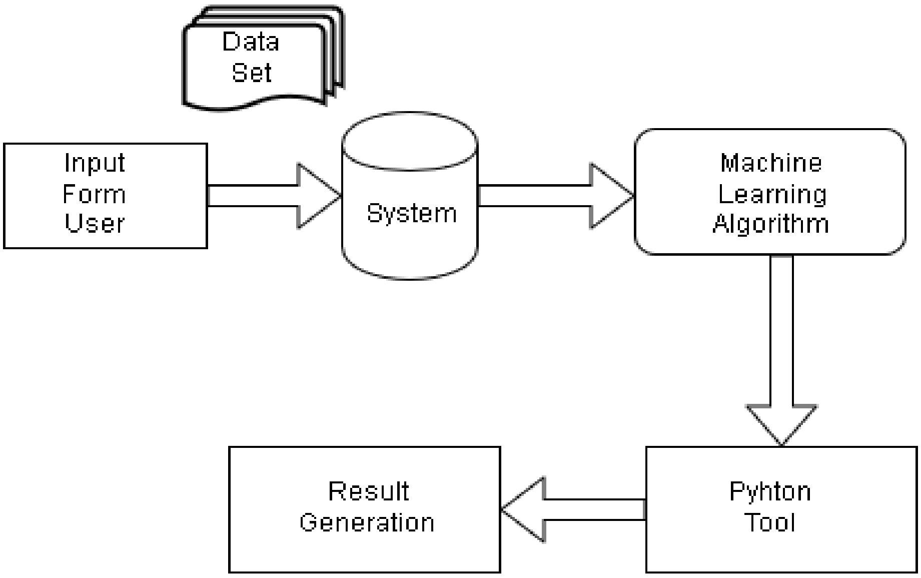


Figure 3.2.3: Flow Chart

3.3 Methodology

3.3.1 Data collection and preprocessing

Data collection is a major bottleneck in machine learning and an active research topic in multiple communities. Data collection largely consists of data acquisition, data labeling, data cleaning and improvement of existing data or models. For a good machine learning project we need a perfect dataset to run our algorithm and perfect dataset generally doesn’t exists and for a project with inappropriate and inadequate data set machine learning gives poor results resulting in less accuracy. So we need a perfect mind setup what kind of data is actually needed for our project.

Data was collected from various medical sites that consists of the dataset of the patients suffering from various diseases. Data which was taken as a reference for our project consists of the fundamental symptoms that caused the disease and many medical tests. Reference data set were taken from UCI Machine Learning Repository and Kaggle which is an online community of data scientists and machine learners, owned by Google, Inc. which allows users to find and publish data sets, explore and build models in a web-based data-science environment. Collected data were then cleaned and normalized for further processing.

3.3.2 Naive Bayes Algorithm:

The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically. Naive Bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method. The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability. Naive bases is often described using categorical data because it is easy to describe and calculate using ratios. A more useful version of the algorithm for our purposes supports numeric attributes and assumes the values of each numerical attribute are normally distributed (fall somewhere on a bell curve). Again, this is a strong assumption, but still gives robust results.

Applications of Naïve Bayes Classification:

- Naïve Bayes Text Classification
- Disease Prediction along with several other algorithms including mapping algorithm, random forest and decision tree
- Spam Filtering
- Recommendation System
- Online applications like Emotion Modelling.

Theory:

R and L are conditionally independent given M if for all x,y,z in {T,F}: P(R=x ½ M=y ^ L=z) = P(R=x ½ M=y). More generally:

Let S1 and S2 and S3 be sets of variables. Set-of-variables S1 and set-of-variables S2 are conditionally independent given S3 if for all assignments of values to the variables in the sets, P(S1's assignments ½ S2's assignments & S3's assignments) = P(S1's assignments ½ S3's assignments)

P(A|B) = P(A ^B)/P(B)

Therefore P(A^B) = P(A|B).P(B) – also known as Chain Rule

Also P(A^B) = P(B|A).P(A)
Therefore P(A|B) = P(B|A).P(A)/P(B)

P(A,B|C) = P(A^B^C)/P(C)

= P(A|B,C).P(B^C)/P(C) – *applying chain rule*

= P(A|B,C).P(B|C)

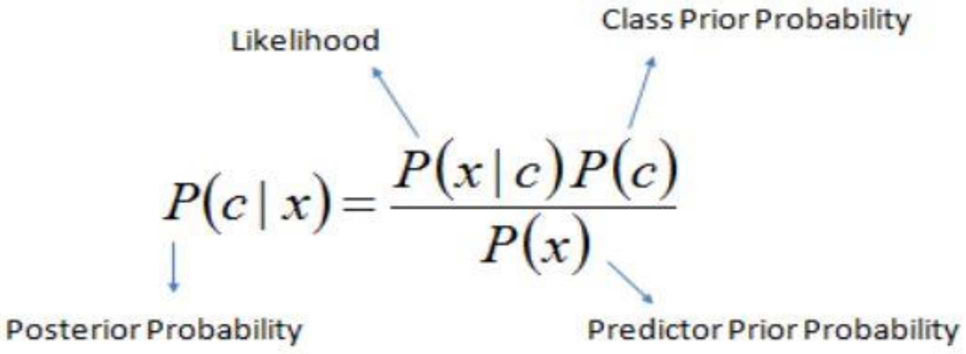
= P(A|C).P(B|C) , *If A and B are conditionally independent given C.*

This can be extended for n values as P(A1,A2…An|C) = P(A1|C).P(A2|C)…P(An|C) *if A1,A2…An are conditionally independent given C.*

$$P(C \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C)P(C)}{P(\mathbf{X})}$$

$$\text{Posterior Probability} = \frac{\text{Likelihood} \times \text{Class Prior Probability}}{\text{Predictior Prior Probability}}$$

Figure 3.3.2.1: Posterior Probability Formula


Likelihood
Class Prior Probability
$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 3.3.2.2: Posterior Probability Formula Labelling

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.

- *P*(*x*|*c*) is the likelihood which is the probability of *predictor* given *class*.
- *P*(*x*) is the prior probability of *predictor*.

Calculation of Mean and Standard deviation in a Normally Distributed Dataset:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{Mean}$$

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2\right]^{0.5} \qquad \text{Standard deviation}$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{Normal distribution}$$

Figure 3.3.2.3: Mean, Standard Deviation and Normal Distribution

Condition Probability modelled with normal distribution:

$$\hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi\sigma_{ji}}}\exp\left(-\frac{(X_j-\mu_{ji})^2}{2\sigma_{ji}^2}\right)$$

Figure 3.3.2.4: Conditional Probability Formula

3.3.3 Algorithm

Step 1: Handle Data: Load the data from CSV file and split it into training and test datasets (Given Split Ratio).

Step 2: Summarize Data: Summarize the properties in the training dataset so that we can calculate probabilities and make predictions which includes Separate Data By Class, Calculate Mean, Calculate Standard Deviation, Summarize Dataset, Summarize attributes by class.

Step 3: Make a Prediction: Use the summaries of the dataset to generate a single prediction which involves the following steps:

- Calculate Gaussian Probability Density Function
- Calculate Class Probabilities
- Make a Prediction
- Estimate Accuracy

Step 4: Make Predictions: Generate predictions given a test dataset and a summarized training dataset. The data from the users are evaluated here, which will generate the necessary results and accuracy.

Step 5: Evaluate Accuracy: Evaluate the accuracy of predictions made for a test dataset as the percentage correct out of all predictions made.

Step 6: Tie it together: Use all of the code elements to present a complete and standalone implementation of the Naive Bayes algorithm.

3.3.4 Data Visualization

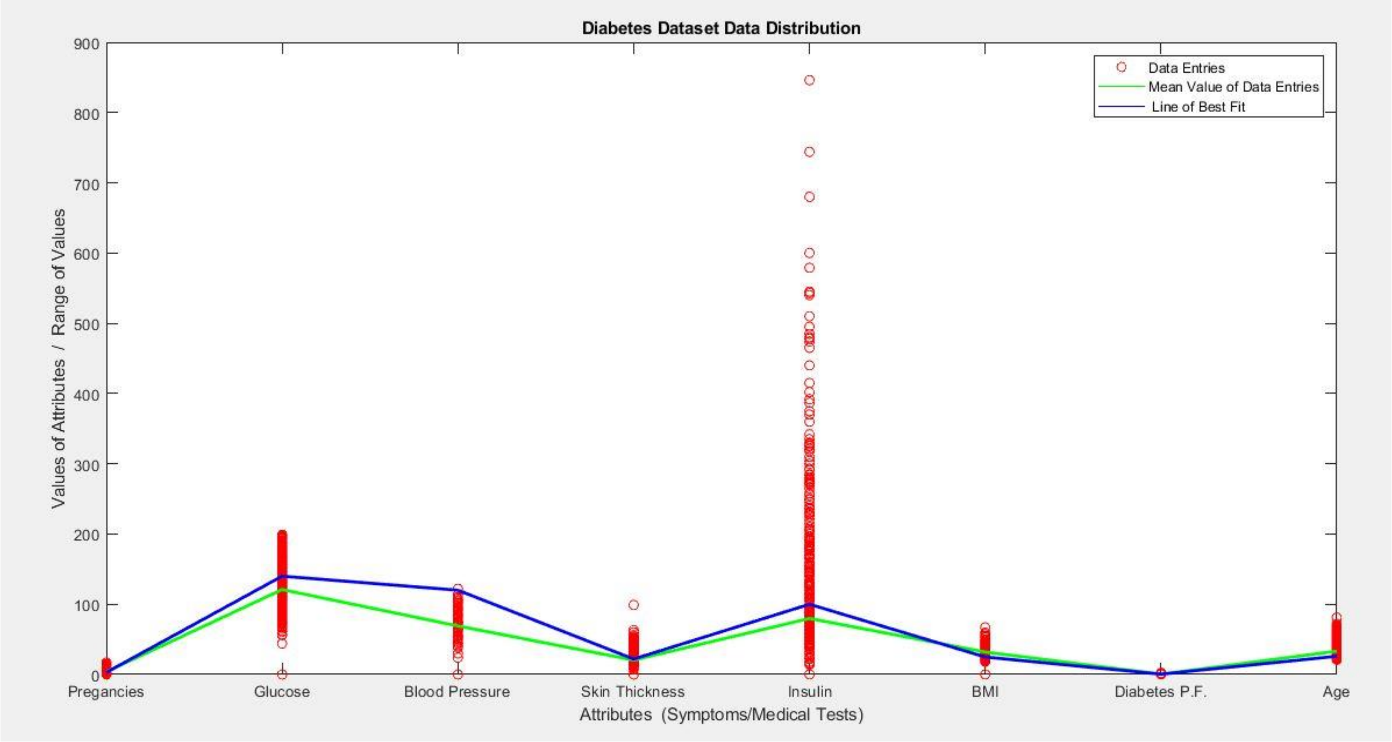


Figure 3.3.1.1: Diabetes Data Visualization using Matlab

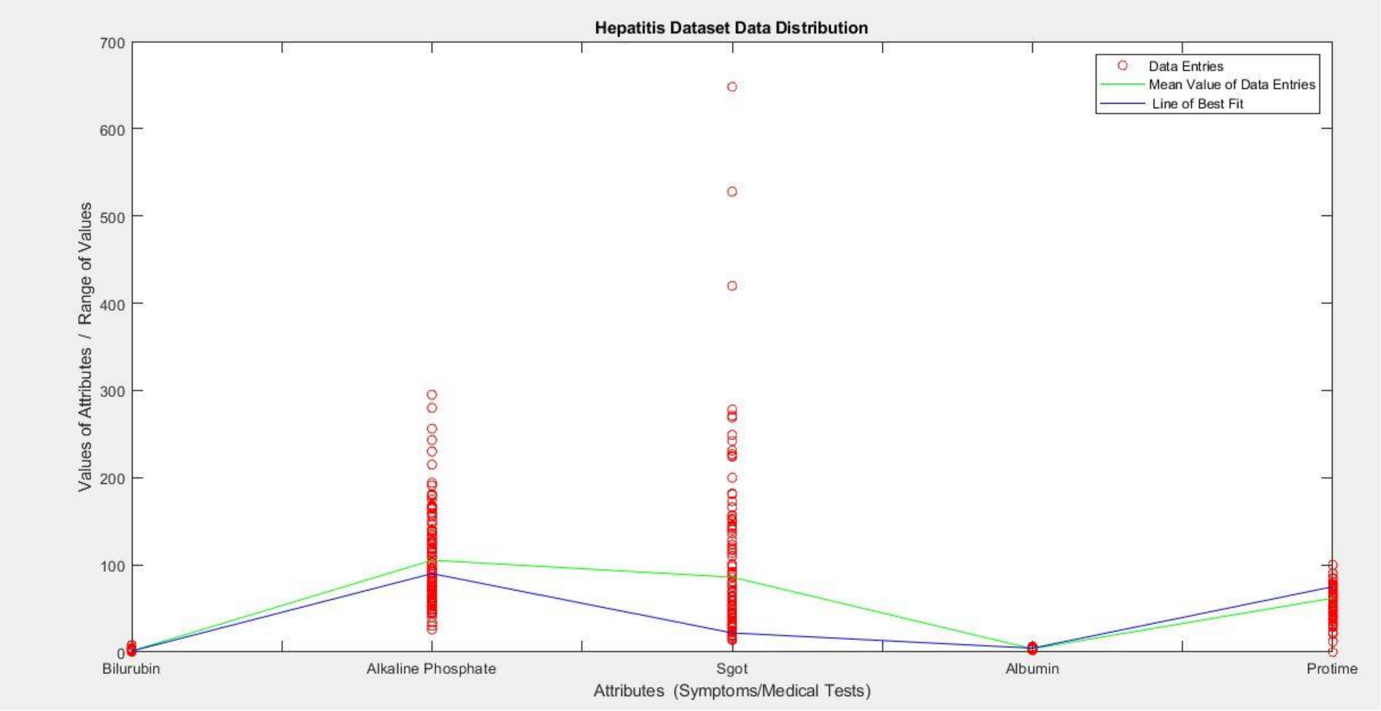


Figure 3.3.4.2: Hepatitis Data Visualization using Matlab

3.3.5 Result

Accuracy for Hepatitis is calculated to be 77.41935% (+/- 2%)by splitting the dataset into 70-30 split ratio (ideal 82% with 10 folds cross validation, 80% with Splitting the training ratio in 70-30 ratio calculated using WEKA tool for the same dataset).

Accuracy for Diabetes is calculated to be 76.30283%(+/- 2%)by splitting the dataset into 70-30 split ratio (ideal 77.47% accuracy using the best performing algorithm "Logistic Regression" and 81.5% accuracy using WEKA tool experimenter)

Other metrics of evaluating the result of prediction:

1. **Confusion Matrix:**

   A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

	Predicted YES	Predicted NO
Actual YES	True Positive (TP)	False Positive(FP)
Actual NO	False Negative(FN)	True Negative(TN)

Table 3.3.5.1: Confusion Matrix Model

- **True positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.
- **True negatives (TN):** We predicted no, and they don't have the disease.
- **False positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- **False negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

Confusion Matrix for Hepatitis Dataset:

	Predicted YES	Predicted NO
Actual YES	TP=27	FP=5
Actual NO	FN=30	TN=93

Table 3.3.5.2: Confusion Matrix for Hepatitis Dataset

Confusion Matrix for Diabetes Dataset:

	Predicted YES	Predicted NO
Actual YES	TP=421	FP=79
Actual NO	FN=103	TN=165

Table 3.3.5.3: Confusion Matrix for Diabetes Dataset

2. Percentage Accuracy:

It is calculated as:

Accuracy= (TP+TN)/Total

Accuracy for Hepatitis Dataset: (27+93)/155=0.7741935

Accuracy for Diabetes Dataset: (421+165)/768=0.763020

3. Recall:

It is the numerical value representing the total number of successful positive prediction of the given class value to the total number of positive predictions made. It is calculated as: Recall = TP/(TP+FN)

Recall for Hepatitis Dataset: 0.4736

Recall for Diabetes Dataset: 0.8034

4. Precision:

It is the numerical value representing the total number of successful positive predictions of the given class value to the total number of actual positive values of the class. It is calculated as: Precision=TP/(TP+FP)

Precision for Hepatitis Dataset: 0.84375

Precision for Diabetes Dataset: 0.842

5. F1 Score:

It is the numerical value representing the harmonic balance between precision and recall value. It is calculated as: F1 Score= (2*Precision *Recall) / (Precision + Recall)

F1 Score for Hepatitis Dataset: 0.606672

F1 Score for Diabetes Dataset: 0.82224

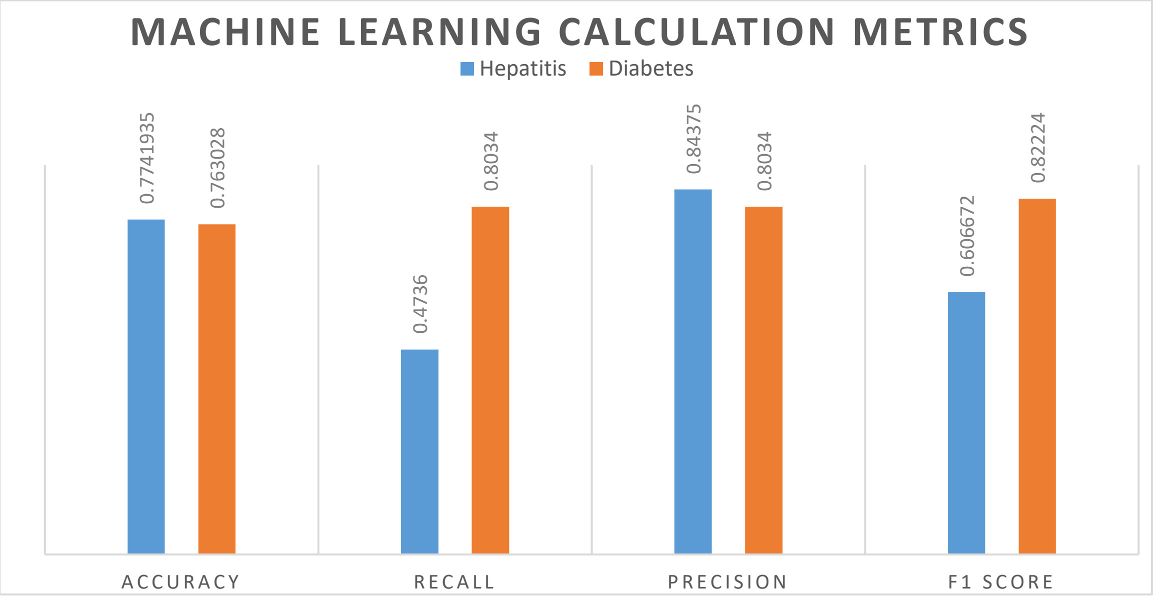


Figure 3.3.5.1: Calculation Metrics

Result Interpretation:

With the accuracy of 77% or 76% approx. in Hepatitis and Diabetes respectively, also based on the symptoms of the individual entry we can successfully predict whether the person might or might not suffer from the given disease with the give accuracy.

For example: For person A with symptoms from 1-9 (attributes), in Hepatitis evaluation, if the result is shown positive for a determining class value, we can say that the person A will be suffering from Hepatitis with 77% accuracy.

Error Analysis:

Errors that might have affected the result:

- Missing data were either deleted (the entire entry) or randomly filled either using the concept of normal distribution, mean and standard deviation or scaling down with mean value compared to the minimum or maximum or simply median value.
- The priority order of attributes based on the probability influence of class value was calculated by preparing a decision tree also using WEKA tool. It was found that the most influencing attributes has a huge variation gaps from the lowest to the highest and a diverse standard deviation from the mean, which effected the overall values.
- Insufficient amount of data for proper training set
- Inconsistency in data

Improving the result:

The result for each dataset can be significantly increased by:

- Use of algorithms for finding out the missing values in big data calculations like libraries from Random Forest, Liner Regression or KNN Algorithm(using the principle of distance measure)

- Proper Data Collection with consistency and managed attributes characteristics
- Large number of data entries for proper training and probability evaluation
- Using X folds Cross Validation of dataset rather than split ratio
- Use of ensemble methods like boosting.

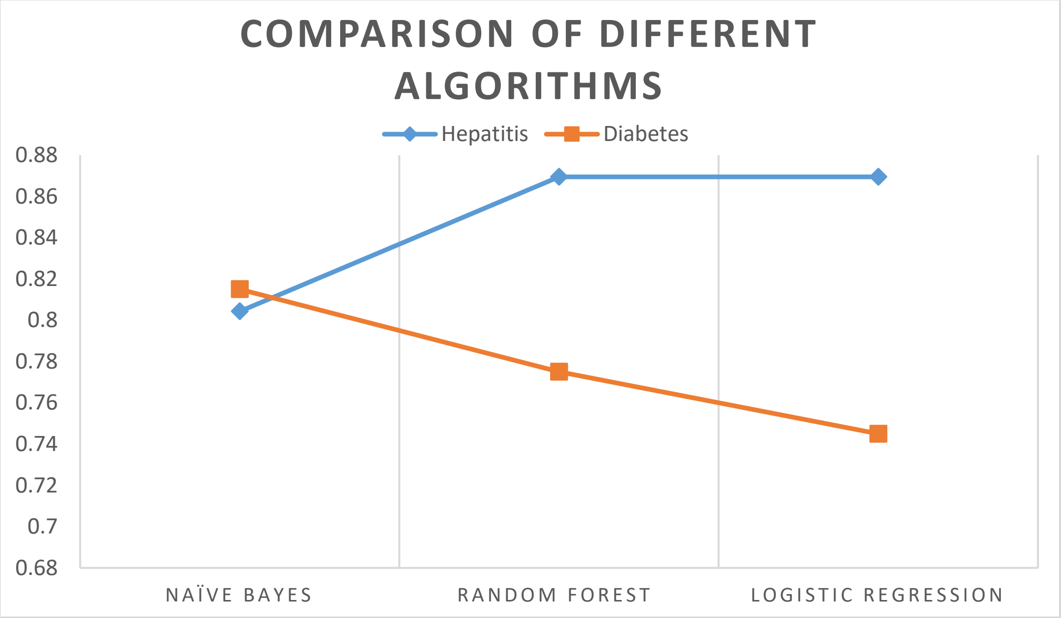


Figure 3.3.5.2: Comparison of Result

All the accuracy calculation is based on the results evaluated in WEKA tool

Comparisons and the accuracy in the above chart were made taking references of the projects conducted by several other students in the same dataset we used with our own accuracy, which clearly shows us that the use of Naïve Bayes algorithm is consistent although there are other algorithms with high accuracy value.

Chapter 4: Discussion on the Achievements

On the completion of our project in machine learning for computer aided diagnosis using Naïve Bayes algorithm for disease prediction we were able to successfully predict whether a person will be or not be suffering from the disease with 75-80% accuracy which will successfully be a head start for further diagnosis and tests to be carried out. In comparing our prediction results with predictions from other algorithms,78% in Two Class Logistic Regression with 10 folds cross validataion,79%+/- 7% using Keras (Kaggle Forum, 2019), and results from other students working in the same dataset, our result is quite good (77% ~78%) overlooking the number of entries and data collection.

(Kaggle Forum, 2019)

Project	**Algorithms**	**% Accuracy**
	Naïve Bayes with 10 folds Cross Validation	81.02%(+/-2%)
	Keras	78.52%(+/-7.49%)
	Two Class Logistic Regression	78.00%(+/-5.009%)
References	SVC	76.43%(+/-5.696%)
	Random Forest Classifier	77.21%(+/-7.129%)
	LDA	77.35%(+/-5.159%)
Self-Diagnosis	Naïve Bayes Algorithm	76.30%(+/-3%)

Table 4.1. Accuracy Comparison (Self Diagnosis vs References)

Chapter 5: Conclusion and Recommendation

5.1 Limitations

The limitations of this project are:

- Datasets used were noisy (diverse) with less entries. (Kaggle, 2018)
- Prediction of missing values using different algorithms wasn't carried out properly due to limited time. (Hepatitis Data Set, 2018)
- Lack of proper hardware, to run simulations in a large number of data

5.2 Future Enhancements

This project can be further enhanced in future by following ways:

- Proper data collection can be done to increase the accuracy of the result by collaborating with medical data centers.
- Prediction can be carried out for numerous diseases based on other regions.
- Health recommendations can be provided to users based on the predicted result.

References

1. Arun.R, M. N. (2018). HEART DISEASE PREDICTION SYSTEM USING NAIVE BAYES. *International Journal of Pure and Applied Mathematics*, 3053-3065.

2. Bharat Gandhi, L. S. (2017, April). Disease Predictor: A Disease Prediction App. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 5*(IV), 1509-1512. Retrieved from https://www.ijraset.com/fileserve.php?FID=7434

3. *Hepatitis Data Set*. (2018, November 1). Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/hepatitis

4. *IDF SEA members*. (2018, November 28). Retrieved from International Diabetes Federation: https://www.idf.org/our-network/regions-members/south-east-asia/members/97-nepal.html

5. Kaggle. (2018, November 20). *Pima Indians Diabetes Database*. Retrieved from Kaggle: https://www.kaggle.com/uciml/pima-indians-diabetes-database

6. *Kaggle Forum*. (2019, 12 20). Retrieved from Kaggle: https://www.kaggle.com/general/19387

7. Mehdi Teimouri, F. F.-M.-D. (2016). Detecting Diseases in Medical Prescriptions Using Data Mining Tools and Combining Techniques. *Iranian Journal of Pharmaceutical Research*, 113-123.

8. Rashid, M. T. (2016, January). Article: Automated Disease Prediction System (ADPS): A User Input-based Reliable Architecture for Disease Prediction. *International Journal of Computer Applications*, 24-29.

9. Samiti, R. S. (2016, April 6). *Nepal at high risk of diabetes*. Retrieved from The HImalayan Times: https://thehimalayantimes.com/kathmandu/nepal-high-risk-diabetes/

10. Shratik J. Mishra, A. M. (2018). GDPS - General Disease Prediction System. *International Research Journal of Engineering and Technology (IRJET)*, 3966-3970.

11. Vinitha S, S. S. (2018). DISEASE PREDICTION USING MACHINE LEARNING OVER BIG DATA. *Computer Science & Engineering: An International Journal (CSEIJ)*.